In-depth Analysis of Tandem Mass Spectrometry Data from Disparate Instrument Types*

Robert J. Chalkley‡, Peter R. Baker, Katalin F. Medzihradszky, Aenoch J. Lynn, and A. L. Burlingame

Mass spectrometric analyses of protein digests produce large numbers of fragmentation spectra that are not identified by routine database searching strategies. Some of these spectra could be identified by development of improved search engines. However, many of these spectra represent fragmentation of peptide components bearing modifications that are not routinely considered in database searches. Here we present new software within Protein Prospector that allows comprehensive analysis of data sets by analyzing the data at increasing levels of depth. Analysis of published data sets is presented to illustrate that the software is not biased to any instrument types. The results show that these data sets contain many modified peptides. As well as searching for known modification types, Protein Prospector permits the detection and identification of unexpected or novel modifications by searching for any mass shift within a user-specified mass range to any chosen amino acid(s). Several modifications never previously reported in proteomics data were identified in these standard data sets using this mass modification searching approach. ***Molecular & Cellular Proteomics 7:2386–2398, 2008.***

There are many search engines available to the researcher for the analysis of proteomics data produced by tandem mass spectrometry. Here we present the performance of one of these: Protein Prospector. The large volume of data typically produced in modern proteomics experiments is too vast for manual interpretation and verification of results, so computer algorithms have been developed for database searching of data and are now routinely relied upon to produce trustworthy results (1). For this approach to be acceptable it is important that a metric of reliability be attached to any peptide or protein identification reported from use of such a database search engine. The most commonly reported measure of this is an expectation value. This value can be calculated by two different approaches. Either a theoretical model is constructed for peptide fragmentation from which a probability and an expectation value can be derived (the approach used by, for example, the Mascot search engine (2)), or search engine-derived matches to a given spectrum deemed to be incorrect are modeled to a distribution, and then a probability and expectation value are derived from this model (the approach used by, for example, X!Tandem (3)). Protein Prospector uses the latter of these two approaches. We used Protein Prospector to analyze a group of "standard" data sets acquired on a variety of different instrument platforms (4) to assess its reliability, sensitivity, and flexibility in analyzing different LC-MS data types using an accepted method to calculate peptide false positive identification rates for the results (5).

Most mass spectrometry search engines are reliable at matching peptide sequences to a significant number of tandem mass spectra. However, in all data sets there are still a large number of spectra that are not matched successfully by search engines. One reason for this situation is that a significant number of the spectra actually correspond to modified peptides. Typical search engine analysis strategies constrain themselves to a very limited number of commonly occurring modifications. However, because some 500 different peptide modifications are already listed in Unimod, use of this type of restricted searching will never extract all the useful information from a data set (21).

The number of modified peptides in proteomic samples is predicted to be very high; for example, a recent estimate suggested the presence of 8–12 modified versions for each unmodified peptide present (6), although most of these modified species are presumed to be present at very low stoichiometry. Several software tools have already been developed to try to identify these modified peptides. Although most conventional search engines can look for a wide variety of modifications using a defined list of choices, they generally suffer from two problems. First the discriminatory power (the ability to distinguish between correct and incorrect answers) drops dramatically when looking for a large number and variety of modifications. This means that, depending on where the threshold is drawn for reporting matches as significant, either fewer answers are reported than in the equivalent search where the modifications were not considered (*i.e.* more false negatives where the search engine gets the correct answer but the match is not deemed significant), or there are

From the Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94158-2517

Received, January 16, 2008, and in revised form, July 14, 2008
Published, MCP Papers in Press, July 24, 2008, DOI 10.1074/mcp.M800021-MCP200

a larger number of false positive identifications. The second issue encountered is that the search speeds become impractically slow because of the increased number of permutations (search space) that need to be considered.

One approach that has been used to tackle the loss of discriminatory power is to tabulate all the mass shifts reported by the software for each amino acid in a large data set (7). Assuming that incorrect assignments will distribute randomly with respect to mass shift and amino acid to which it is assigned, those matches containing mass modifications to particular amino acids that are reported multiple times are more likely to represent *bona fide* modified peptides. This approach works for identifying mass modifications for the peptide but is often unreliable at determining the correct residue of modification. This is because the mass spectra that are being searched often do not contain sufficient information to distinguish the exact residue of modification. Hence for example, a mass shift of +16 Da (oxidation) is reported more often than at random for practically all amino acids, even those that cannot become oxidized.

The second issue of addressing the search speed can be handled by looking for the modifications on only a small subset of all peptides/proteins in the database. This is commonly done by first searching the data set using a conventional search engine strategy focusing on reliably identifying proteins in the sample on the basis of unmodified peptides. Then only those proteins (7) or peptides (8) identified in the initial search are considered in the search for modified peptides. This assumes that modified peptides will only be found associated with proteins for which unmodified peptides were also present. This type of modification search is sometimes done by trying to correlate the spacing of peaks in unidentified spectra to those in spectra identified in the initial search (7). Alternatively *de novo* interpretation of some or all of the amino acid sequence (9, 10) can be used to restrict the search space.

Here we describe the adaption of an existing algorithm, Batch-Tag, to allow searching of tandem mass spectra for mass modifications. The search can consider any mass modification within a mass range (positive or negative modification) on either selected or all amino acid residues on the N or C terminus or a modification that is observed as a neutral loss in MSMS analysis (preventing modification site assignment) (11, 12). Like other software, these searches are normally performed against a particular list of database protein accession numbers that were identified in an initial search (that considered only a restricted list of modifications).

MATERIALS AND METHODS

All the data used in this study were created in a previously published study, so for experimental details see Klimek *et al.* (4). For analysis of data acquired on different instruments, mzXML peak lists posted on the Seattle Proteome Center Public Data Repository were used (regis-web.systemsbiology.net/PublicDatasets). The four data files analyzed were all from Mix 2 and were named QS20060131_S_18mix_02, LT20060105_S_18mix_03, run1pps_D06_03005_hlee_18pro_mix_1144399822, and C0605_000118.

Depending on the instrument setting chosen, Protein Prospector may attempt to determine fragment ion charges and deisotope the peak list: it will deisotope 4800 data and deisotope and perform charge state determination (for 1+ and 2+ fragments based on the isotope spacing) for QSTAR data but will not attempt these preprocessing steps for low resolution ion trap data. Protein Prospector then splits the mass range covered by the fragment ions in the peak list in half and uses the 20 most abundant peaks in each half of the spectrum for database searching (13).

All searches were performed against a Swiss-Prot database downloaded on May 24, 2007 to which a randomized version of the same database was concatenated to give a total of 534,708 protein entries. For initial searches, full tryptic specificity was required allowing for one missed cleavage. Oxidation of methionine, protein N-terminal acetylation, carbamidomethylation of cysteines, and pyroglutamate formation from N-terminal glutamine residues were the only considered variable modifications, and up to two of these were allowed per peptide. For the extensive searches the list of protein accession numbers identified in the initial search for each instrument was used, and nonspecific cleavages at both ends of the peptides were considered (similar to a no enzyme specificity search except the one missed trypsin cleavage requirement was maintained). Variable modifications considered were acetyl (Lys), acetyl (protein N terminus), acetyl + oxidation (protein N-terminal Met), Asn → succinimide (Asn), carbamidomethyl (Cys), carbamyl (Lys), carbamyl (N terminus), deamidated (Asn), dethiomethyl (Met), dioxidation (Met), Gln → pyro-Glu (N-terminal Gln), Glu → pyro-Glu (N-terminal Glu), Gly-Gly (Lys) (*i.e.* ubiquitination), HexNAc (Ser or Thr), Met loss (protein N-terminal Met), Met loss + acetyl (protein N-terminal Met), methyl (Lys), nitro (Tyr), oxidation (Met), oxidation (Trp), phospho (Ser, Thr, or Tyr), trioxidation (Cys), Trp → hydroxykynurenine (Trp), and Trp → kynurenine (Trp).

For the XCT data file (C0605_000118), all spectra were considered as 2+ or 3+ precursors. Parent mass tolerance was 2.5 Da; fragment mass tolerance was 0.6 Da. Instrument type was set to ESI-ION-TRAP-low-res. For the LTQ data file (LT20060105_S_18mix_03) all spectra were considered as 2+ or 3+ precursors. Parent mass tolerance was 2 Da; fragment mass tolerance was 0.6 Da. Instrument type was set to ESI-ION-TRAP-low-res. For the QSTAR data file (QS20060131_S_18mix_02) all spectra were considered as 2+ or 3+ precursors. Parent mass tolerance was 100 ppm, and a fragment mass tolerance of 0.1 Da was used. Instrument type was set to ESI-Q-TOF. For the 4800 data file (run1pps_D06_03005_hlee_18pro_mix_1144399822) all spectra were assumed to be singly charged, and parent and fragment mass tolerances of 400 ppm and 0.2 Da were considered. Instrument type was set to MALDI-TOF-TOF. All the mass tolerance parameters were chosen based on cursory examination of the mass accuracy of the data sets. Supplemental Fig. 1 shows histograms of the precursor mass accuracy for all reported assignments. This shows that there was a systematic error in the QSTAR raw data and that the 4800 data were unusually poorly calibrated. For data from all instruments, the acceptance threshold was set as the E-value at which a 2.5% peptide false positive rate was reached according to the target-decoy database search results. The previously published results for these data sets were based on searching using Sequest (14), and they also reported a 2.5% false positive rate.

Protein Prospector calculates expectation values based on random answers. Each spectrum, as well as being searched against the specified database, is also searched against a database of randomized (sequence-shuffled) sequences. This is performed to reduce the chance of homologous peptide sequences being present in the distribution, which is more likely if the results from a normal database

search are used after removing the top answer. Also as the normal database is not random in nature, some sequences occur many times, which can create spikes in the distribution. The results from the randomized database search are plotted as a graph of score *versus* log survival where survival is the ratio of all matches that exceed a given score. From the highest scoring 10% of the distribution a linear fit is calculated. The gradient and offset from this fit are then used to extrapolate probabilities that a given score achieved in the search against the normal target database is part of this incorrect score distribution. To convert the probability to an expectation value, the score probability is multiplied by the number of peptides in the normal database that fit the precursor ion requirements; *i.e.* the peptide is predicted to be formed by the specified enzyme cleavage and has the correct mass (within the mass tolerance of the search). This method for calculating the probability of a score being part of the random distribution is similar to that proposed by Fenyo and Beavis (15) except with the Protein Prospector scoring system the plot of score rather than log score is more linear.

For mass modification searching of QSTAR data, the mzXML peak list from file QS20060131_S_18mix_02 was modified so that a doubly and triply charged version of every peak list was present (to remove variability due to how the different softwares decide on precursor charge state). Using Protein Prospector, the same four variable modifications considered in the initial searches (oxidation (Met), Gln → pyro-Glu, carbamidomethyl (Cys), and protein N-terminal acetylation) were considered with up to two modifications per peptide. In addition, a single mass modification between −100 and +300 Da was considered on any amino acid. The mass modification range is specified as integers, but a mass defect is applied to the modification during the search. This is because a modification is never an exact integer mass, so if a mass defect is not applied this will introduce a significant loss in mass accuracy to assignments, requiring an opening of precursor and fragment mass tolerances considered to be able to assign the modified peptides. Applying a mass defect (the default is to add 0.00048 Da per amu) allows data to still be searched without seriously compromising the level of mass accuracy tolerance even when the structure of the modification is not known. The benefit of using the mass defect is more pronounced for higher mass accuracy data. It is also increasingly important as you move to larger mass modifications; *e.g.* for a 300-Da modification this corresponds to a mass shift of 0.144 Da. For the mass modification searches of QSTAR data described in this study it reduces the number of precursors considered by about a third compared with searching with 0.5-Da mass accuracy on the precursor, resulting in a similar level effect on search speed and false positive matches.

The requirement for tryptic specificity was removed at both termini. Parent mass tolerance was set to 0.25 Da, and other parameters were the same as in the previous QSTAR searches. In all other respects, this searching is identical to a search with defined modifications, *i.e.* comparing theoretical masses with observed masses without any other filtering. Hence this is covering a very large search space, and the restriction to searching only a short list of proteins is necessary if large numbers (thousands) of spectral peak lists are to be searched. This particular search of just under 4000 peak lists searched against 46 proteins (including homologs) completes in about 8 h on a dual processor 2.8-GHz Intel Xeon desktop computer.

For Inspect (version downloaded July 2007), instrument type was set as Q-TOF, protease was set as trypsin, and a "blind" search was performed allowing one mass modification. Parent tolerance was set to 0.25 Da, and fragment tolerance was set to 0.1 Da. A maximum post-translational modification size of 300 Da was specified.

If search parameters are changed in a way that will alter peptide score distributions (*e.g.* considering more modifications), then the random score distributions should be redetermined so accurate probabilities can be reported and used for expectation value calculation. Also allowing for more modifications will increase the number of potential peptides with the correct precursor mass, so the conversion value from probability to expectation value will also change. Hence if one compares a result from a data set searched using a list of defined mass modifications with the result of the same peak list searched allowing for undefined mass modifications, the same match will have the same peptide score (based on number of b, y… ions matched) but different expectation values. The searching of a restricted number of database entries allowing for a vast range of modifications (*i.e.* a mass modification search) creates complications in calculating an accurate expectation value. Probabilities for a given score should be reasonably accurate. However, the number of entries with the correct precursor mass in this type of search (used to convert the probability to an expectation value) could be an inaccurate measure. On the one hand a very small protein database is being used, but conversely if you allow for mass modifications of −100 to +300 Da this will mean that any peptide in the database whose unmodified mass is within this range of the precursor ion will be considered, so these two factors balance each other out to some extent. Nevertheless any inaccuracy in the expectation value measure will be the same to matches to a decoy database. When Protein Prospector searches a concatenated normal-randomized database when a list of accession numbers has been specified, it will search against the randomized versions of the same protein entries. Hence the false discovery rate estimation should still be accurate.

RESULTS

Comparison of Peptide Identifications on Different Types of MSMS Data—To assess performance of a search engine it is useful to perform analyses of standard data sets that are available in the public domain. One of the largest and most versatile set of standard data sets is of a mixture of nominally 18 standard proteins that has been run on a wide variety of different instruments (4).

For this assessment of Protein Prospector we chose to analyze data sets from four different instruments used for Mix 2 analysis: an XCT (three-dimensional ion trap from Agilent/Bruker), LTQ (linear ion trap from Thermo), QSTAR (ESI-Q-TOF instrument from Sciex/Applied Biosystems), and a 4800 (MALDI-TOF-TOF (with a quadrupole collision cell between TOFs) from Applied Biosystems), which together represent the major types of CID fragmentation data in use. First these published data sets were analyzed using standard search parameters (assuming fully tryptic cleavages and only minimal modifications allowed). Using stringent acceptance criteria, a list of peptides and proteins was acquired for each sample. This list of proteins was then used to restrict the database entries during further searches of the data. In this second level analysis, semi- and completely non-tryptic peptides were also considered as well as a very wide range of defined potential modifications. This search is analogous to the Sequest searches performed in the publication accompanying these data sets (4). The numbers of peptides identified in these secondary analyses are presented in Table I and plotted as receiver-operating characteristic curves in Fig. 1, and the full search results are in supplemental Table 1. Also plotted in Fig. 1 are the single data point results reported for the Sequest searches (4).

It is possible to make comparisons of the results presented here with the previously published search results using Sequest. However, exact number comparisons are not meaningful as different search parameters were used, and the numbers presented in the study are averages for 10 replicate data sets, whereas the Prospector results are just one example of each data set. The salient observation from these results is the change in total number of matches at a given rate of false positive identifications (*e.g.* the points at which the *curves* intersect the *dotted line* in Fig. 1). The graph shows that the results for XCT data are very similar from the two search engines but that Protein Prospector identified significantly higher numbers of components at the same false discovery rate threshold for the other three instruments. Also whereas the Sequest results suggested that the LTQ performed dramatically better than other platforms, the Protein Prospector results indicate much smaller differences between LTQ, QSTAR, and 4800 data set results.

As can be seen from supplemental Table 1 and also noted in the previous analysis of these data sets (4) a large number of semi- and non-tryptic peptides are present as well as many peptides observed with modifications that are formed when samples are stored for long periods, such as deamidation, oxidation, and succinimide formation from asparagine residues. This is not unexpected for a sample mixture that is derived from commercial protein sources that will have been purified long before the sample was analyzed. Because of these factors, this sample mixture is a reasonable test bed for evaluating database search software designed to find unexpected modifications.

Identifying Unexpected and Novel Modifications—MS-Alignment, part of the Inspect software package, is a leading freely available software tool that assigns unpredicted mass modifications (7, 16). Therefore, MS-Alignment was used in this study to evaluate the performance of Protein Prospector in finding unexpected mass modifications searching the QSTAR data set analyzed above, and then the results from MS-Alignment were compared with analysis using Protein Prospector. Assigning modifications to residues is a much less reliable process than identifying peptides. The problem stems from the difficulty in differentiating between a result that is homologous to the correct answer and the actual correct answer. For modification analysis, a search engine can generally reliably assign a spectrum to a given peptide sequence in the database with the correct modification mass, but reporting of the site of modification is often not reliable partly because there is frequently insufficient information to determine the exact residue of modification. When searching for mass modifications of undefined mass values this problem is exacerbated, and it is common to obtain a peptide match where a large part of the sequence is matched, but the region containing the reported modification is not identified. Matches where the peptide assignment is correct but the modification and site assignment are not have been referred to as delta mass correct (7).

Supplemental Table 2 presents a comparison of spectral assignments by Inspect and Protein Prospector for all peak lists in the QSTAR standard data set examined in the first part of this study. At first glance, the level of correlation of results between the two search engines is low. However, a significant contribution to this apparent discrepancy in assignments between the two search engines are situations where the same

TABLE I

Comparison of peptide identifications using different instruments between Protein Prospector and Sequest

Both results are reported at a 2.5% peptide false discovery rate. The numbers in parentheses in the Protein Prospector (PP) column correspond to the number of decoy matches at this threshold.

Instrument	PP, non-tryptic, defined modifications	Sequest average
QSTAR	1046 (13)	485.6
XCT	623 (7)	604.4
LTQ	1318 (16)	1033.1
4800	1235 (15)	687.8

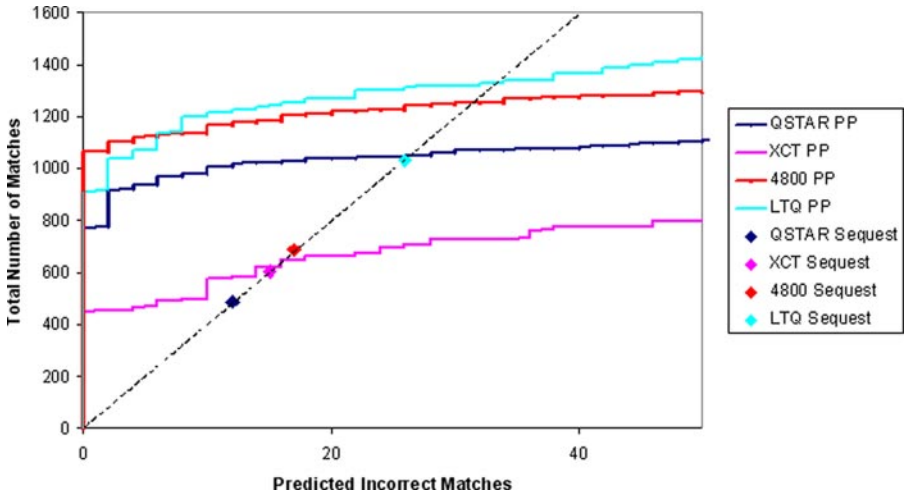


FIG. 1. **Receiver-operating characteristic curves for Protein Prospector (*PP*) searches of the different instrument data.** The predicted number of incorrect answers is derived by doubling the number of matches to the decoy part of the database in the concatenated database search. The data points for the Sequest results are derived from the average number of peptides reported in the publication for each instrument with the number of predicted incorrect results indicated assuming a 2.5% FDR as reported in the publication.

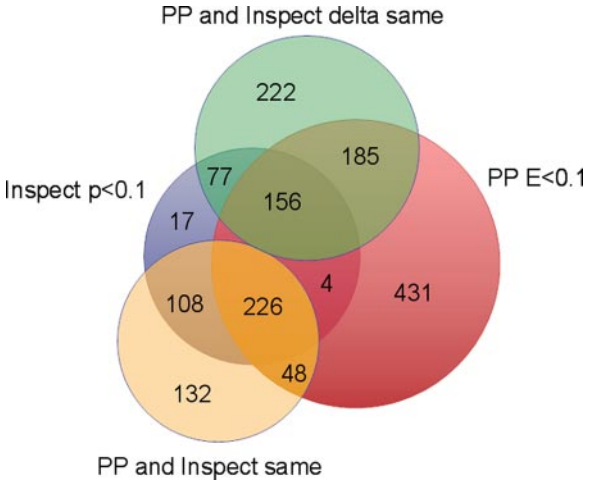


FIG. 2. **Venn diagram showing the levels of overlap between Protein Prospector (*PP*) and Inspect results, both without any confidence level threshold and while applying thresholds of E-value < 0. 1 and $p < 0.1$.**

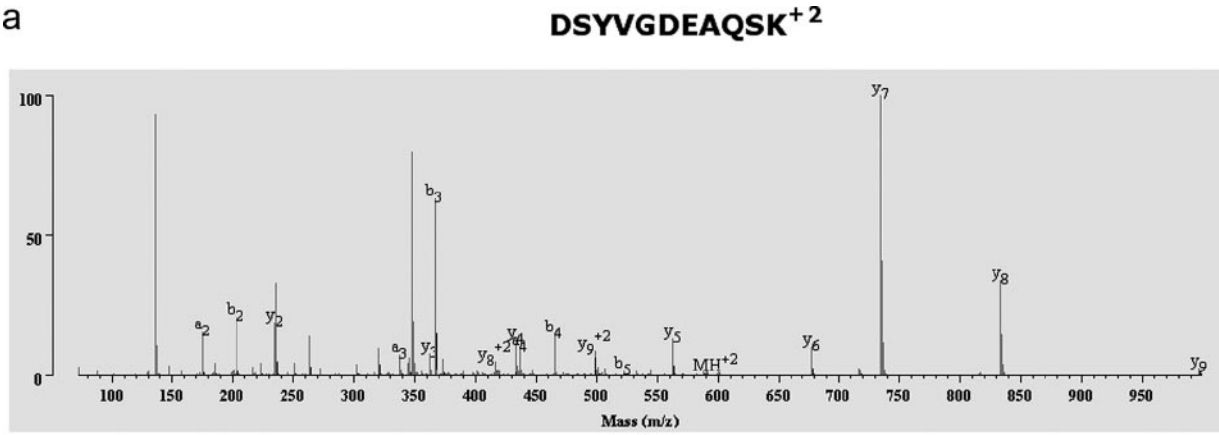
peptide has been reported but the mass modification has been reported differently. For example, one search engine may report a tryptic peptide sequence with an undefined mass modification, whereas the other reports the sequence with an extra amino acid residue at one of the termini that corresponds in mass to the undefined mass modification reported by the other search engine. Unfortunately although these search engines are very powerful at identifying peptides and mass modifications, they currently lack "common sense" so given the opportunity will sometimes report a more convoluted interpretation when a simpler explanation exists. To try to minimize this phenomenon, when several interpretations of a spectrum achieve the same score, Protein Prospector will report assignments that contain no undefined mass modifications in preference to one containing an unnamed mass modification.

Comparing the results of the two search engines, it is clear that Inspect allows a wide degree of variability on the precursor and fragment ion masses; *i.e.* Inspect often reported deamidations when there was no evidence of a mass shift, and it would report peptides as unmodified when the precursor mass was incorrect by several daltons. Thus, for the comparison of results it was decided to ignore these as differences, and the final column in supplemental Table 2 indicates whether matches of the two programs were considered the same (1), delta mass the same (2), or different (0). A summary of the overlap of Protein Prospector and Inspect results is shown in Fig. 2. Comparing the two sets of results, Protein Prospector and Inspect report the same peptide with the same modification and site for 514 (of 3734) of the peak lists. For a further 640 the answers are both peptide and delta mass the same. Both search engines report confidence measures with their assignments. Inspect reports 588 matches with p values of less than 0.1, whereas Protein Prospector reports 1050 matches with E-values less than 0.1. Of the 588 confident Inspect results Protein Prospector agrees according to peptide and at least delta mass on 568 occasions. Conversely Inspect only agrees with 615 of the 1050 confident matches reported by Protein Prospector. This Venn diagram also shows that there are considerable numbers of assignments that both Inspect and Protein Prospector agree upon, but neither assigned a high confidence to the assignment. Many of these are due to wrong charge state assignments, and this is discussed further below.

Of course, it would be preferable to know which matches are being assigned correctly. Unfortunately as these are real data, the correct answers cannot be definitely known, so a certain level of subjectivity is introduced when deciding which answers are believed to be correct. Our impression from looking through the assignments is that Protein Prospector is getting the correct or delta mass correct answer more often than Inspect but that the difference is not as large as the 588 *versus* 1050 number disparity suggests; *i.e.* the Inspect p values are probably more conservative than the Protein Prospector E-values.

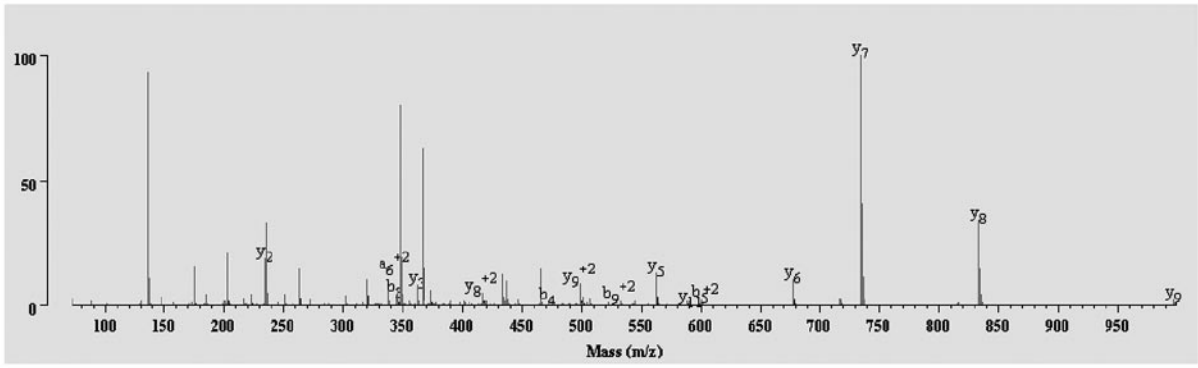
The Protein Prospector search was repeated against a concatenated normal-randomized database to get an estimate of a false positive rate (data not shown). The first random database hit had an expectation value of 0.25, and the second had an expectation value of 0.99. Hence at the 0.1 expectation value threshold used, practically all the results are non-random. This does not necessarily mean they are identical to the correct answer but that they are at least homologous. Indeed as there are only two random database answers with expectation values less than 1, this would suggest a higher acceptance threshold could be used. However, there is one significant caveat with this set of peak lists that has to be considered. As each spectrum is being searched with peak lists representing it as a 2+ and 3+ precursor, there are going to be a number of peak list matches that are to the wrong precursor charge state but still report the same core peptide (but either longer or shorter depending on whether the precursor charge state was higher or lower than that assigned to the peak list). An example of this phenomenon is shown in Fig. 3, which shows the matches to peak lists 501 and 502, representing the 2+ and 3+ versions of one spectrum. This precursor was actually doubly charged, so the result for peak list 501 (DSYVGDEAQSK) (Fig. 3*a*) is the correct assignment for this spectrum. However, peak list 502 is matched to an extended version of this peptide by both Protein Prospector and Inspect (Protein Prospector reports 82.0394-GM(oxidation)GQKDSYVGDEAQSK as shown in Fig. 3*b*). y2–y9 ions are present in the peak list and are the same for the 2+ and 3+ assignments. Four b ions are matched to the 2+ version of the spectrum, whereas three b ions and a doubly charged b ion match to the assignment to the triply charged peak list derived from the same spectrum. Indeed this is not a problem unique to mass modification searching as it still occurs (al-

a



y2+	y		b
		D	
542.2513	1083.4953	S	203.0662
498.7353	996.4633	Y	366.1296
417.2036	833.3999	V	465.1980
367.6694	734.3315	G	522.2195
339.1587	677.3101	D	637.2464
281.6452	562.2831	E	766.2890
217.1239	433.2405	A	837.3261
181.6053	362.2034	Q	965.3847
117.5761	234.1448	S	1052.4167
74.0600	147.1128	K	

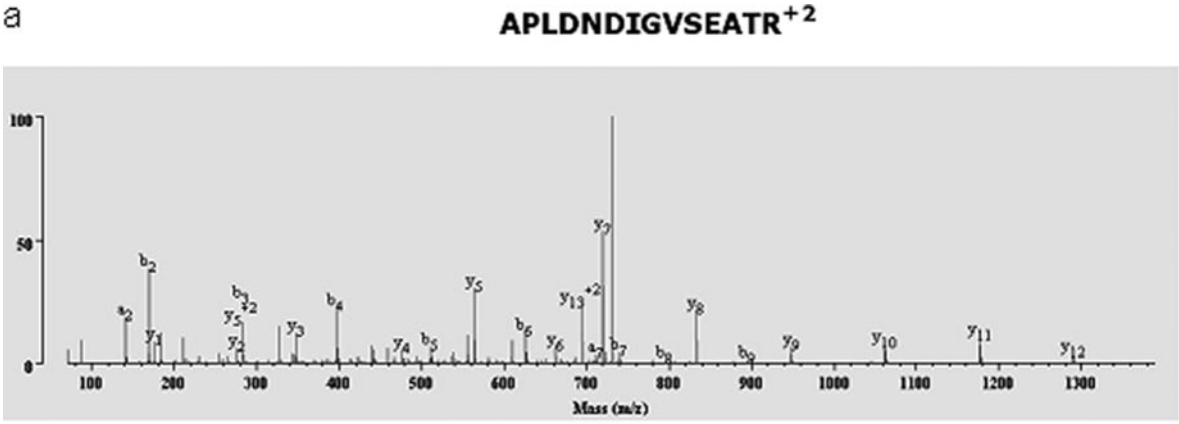
b



y2+	y	82.039-G M(Oxidation)	b	b2+
829.8700	1658.7326		287.1035	
756.3523	1511.6972	G	344.1250	
727.8415	1454.6758	Q	472.1836	
633.8122	1326.6172	K	600.2785	300.6429
599.7648	1198.5222	D	715.3055	358.1564
542.2513	1083.4953	S	802.3375	401.6724
498.7353	996.4633	Y	965.4008	483.2041
417.2036	833.3999	V	1064.4693	532.7383
367.6694	734.3315	G	1121.4907	561.2490
339.1587	677.3101	D	1236.5177	618.7625
281.6452	562.2831	E	1365.5603	683.2838
217.1239	433.2405	A	1436.5974	718.8023
181.6053	362.2034	Q	1564.6559	782.8316
117.5761	234.1448	S	1651.688	826.3476
74.0600	147.1128	K		

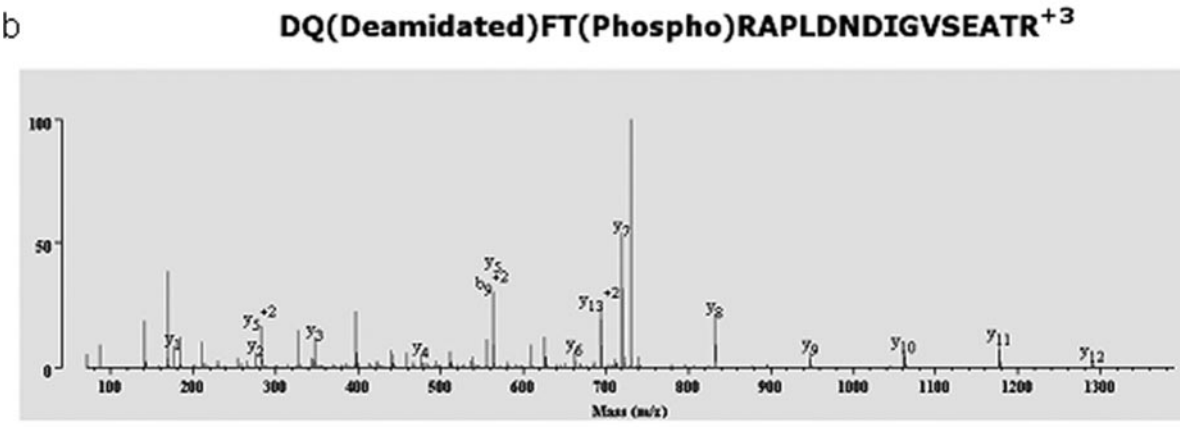
FIG. 3. **Duplication of peak lists for different charge states causes false positive identifications of the wrong charge state peak list to homologous peptides to the correct answer.** Both of the matches presented are derived from the same mass spectrum. The match in *a* is assuming a 2+ precursor charge state; the match in *b* is assuming a 3+ precursor charge state.

a



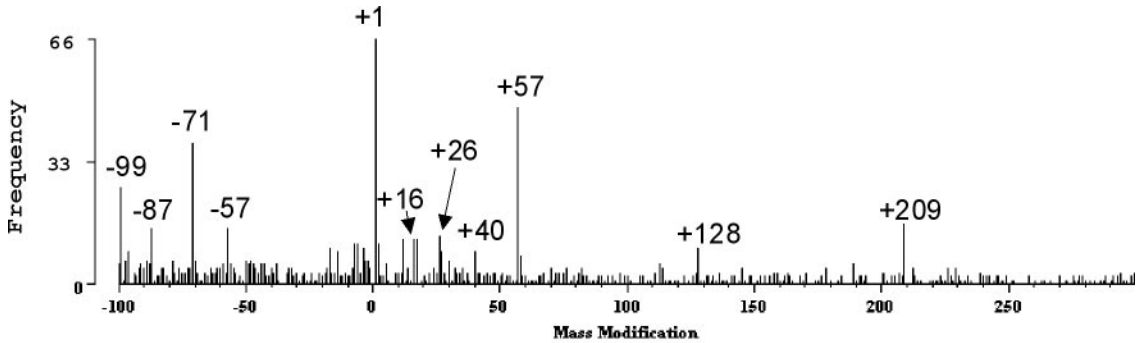
y2+	y		b
		A	
693.8466	1386.6859	P	169.0972
645.3202	1289.6332	L	282.1812
588.7782	1176.5491	D	397.2082
531.2647	1061.5222	N	511.2511
474.2433	947.4793	D	626.2780
416.7298	832.4523	I	739.3621
360.1878	719.3682	G	796.3836
331.6770	662.3468	V	895.4520
282.1428	563.2784	S	982.4840
238.6268	476.2463	E	1111.5266
174.1055	347.2037	A	1182.5637
138.5870	276.1666	T	1283.6114
88.0631	175.1190	R	

b



y2+	y		b	b2+
		D		
1035.9782	2070.9492	Q(Deamidated)	245.0768	
971.4669	1941.9066	F	392.1452	
897.9227	1794.8382	T(Phospho)	573.1592	
807.4157	1613.8242	R	729.2603	365.1338
729.3652	1457.7231	A	800.2975	400.6524
693.8466	1386.6859	P	897.3502	449.1788
645.3202	1289.6332	L	1010.4343	505.7208
588.7782	1176.5491	D	1125.4612	563.2343
531.2647	1061.5222	N	1239.5042	620.2557
474.2433	947.4793	D	1354.5311	677.7692
416.7298	832.4523	I	1467.6152	734.3112
360.1878	719.3682	G	1524.6366	762.8220
331.6770	662.3468	V	1623.7050	812.3562
282.1428	563.2784	S	1710.7371	855.8722
238.6268	476.2463	E	1839.7797	920.3935
174.1055	347.2037	A	1910.8168	955.9120
138.5870	276.1666	T	2011.8645	1006.4359
88.0631	175.1190	R		

FIG. 4. **Different charge state assignments to the same precursor can also lead to false positive matches in searches with defined variable mass modifications.** *a* and *b* show the matches to 2+ and 3+ precursor charge state peak lists of the spectrum that was acquired at 34.1304 min.



-99, -87, -71, -57: Amino acid losses
+1: Combination of de-amidation and wrong precursor isotope.
+16: Oxidation
+26: Schiff base on N-terminus
+40: pyrolyzed carbamidomethyl cysteine
+57: Carbamidomethylation
+128: Lysine Addition
+209: Carbamidomethylated DTT modification

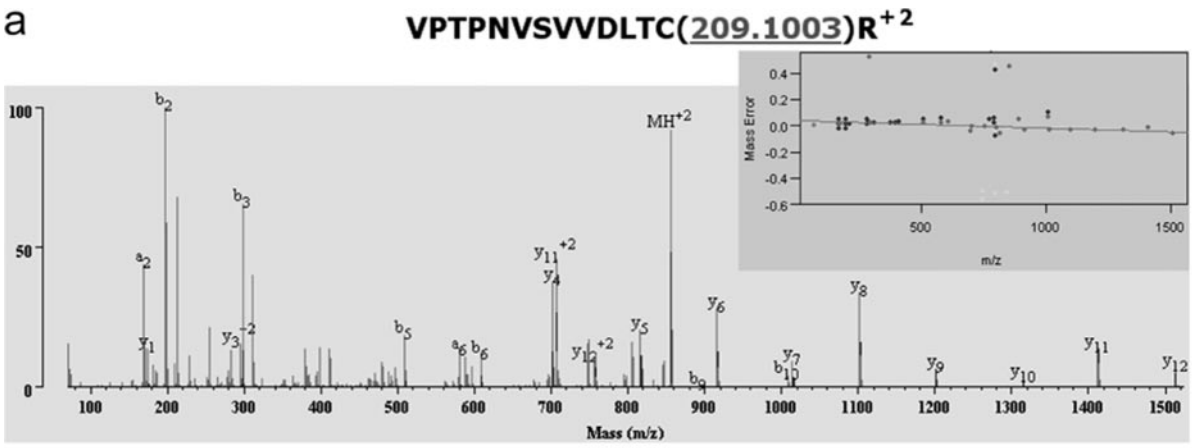
FIG. 5. **Histogram of mass modifications reported by Protein Prospector when analyzing the QSTAR standard data set.**

though much less frequently) in defined modification searches. Fig. 4 shows an example from the QSTAR defined variable modification search for a spectrum acquired at 34.1304 min (reported in supplemental Table 1). Although the match in Fig. 4*a* to the doubly charged peak list is clearly the correct answer, the match of the triply charged peak list of this spectrum in Fig. 4*b* to the extended peptide with a deamidation and phosphorylation is also a significant scoring match. In fact, we believe there are no phosphorylated peptides in this sample because of the presence of alkaline phosphatase, which should have removed all phosphorylations. This duplication of charge states thus makes determination of a sensible threshold for acceptance more complicated, and if the original mzXML peak list had charge states assigned the process would have been much simpler.

Most search engines are poor at representing the level of ambiguity in modification site assignments. Auxiliary softwares have been written to try to address this issue (17, 18) but are not used in these versions of Protein Prospector or Inspect, so for the data presented here, manual verification of site assignments is necessary. However, we argue that results of unknown mass modification searches are probably not sufficiently robust to accept unsupervised anyway, so manual verification would be advisable even if such a site assignment score were reported.

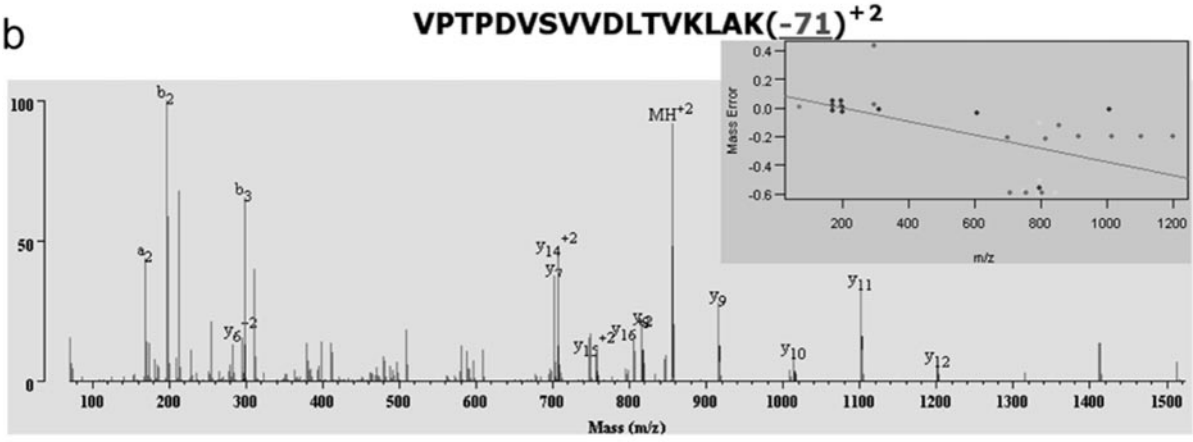
Fig. 5 shows a histogram of all the mass modifications reported in the Protein Prospector search along with explanation for the major peaks. Several of the major peaks correspond to amino acid losses. This is caused because the addition of an extra amino acid to the sequence followed by reporting this mass as a loss gives essentially the same list of potential fragments with the exception of the addition of an extra potential b and y ion (because of the increased length), meaning this can get a larger score than the unmodified equivalent. This problem is caused by the combination of performing mass modification searching with removing any required enzyme specificity for cleavage. Modifications of +16 Da and +57 Da correspond to oxidation and carbamidomethylation. These peaks are disproportionately small in this histogram because these two modifications were both considered as defined modifications to methionine and cysteine, respectively, so if they were matched to the specific modification on the correct residue then they were not reported as a mass modification and so are not represented in this histogram. A number of identifications are reported as peptides with modifications of ±2–8 daltons of which there were about 20 confident identifications. What these actually represent are spectra where a second, generally more intense, precursor ion was co-isolated and fragmented at the same time as the targeted precursor ion. The other major group of new identifications is peptides containing amino acid substitutions. Upon searching the literature, many of these substitutions have been reported previously. Some peptides with unusual mass modifications that are not listed in modification repositories such as Unimod were also identified. 18 spectra were reported by Protein Prospector with a mass modification of nominally 209 Da to cysteine residues of which 14 had expectation values of less than 0.1. For example, Fig. 6*a* shows spectrum 1975, which was matched to VPTPNVSVVDLTC(209)R. The matching of y1 and y4–y12 ions gives a confident match and suggests the modification of 209 Da is on the cysteine residue. Inspect reported a different

a

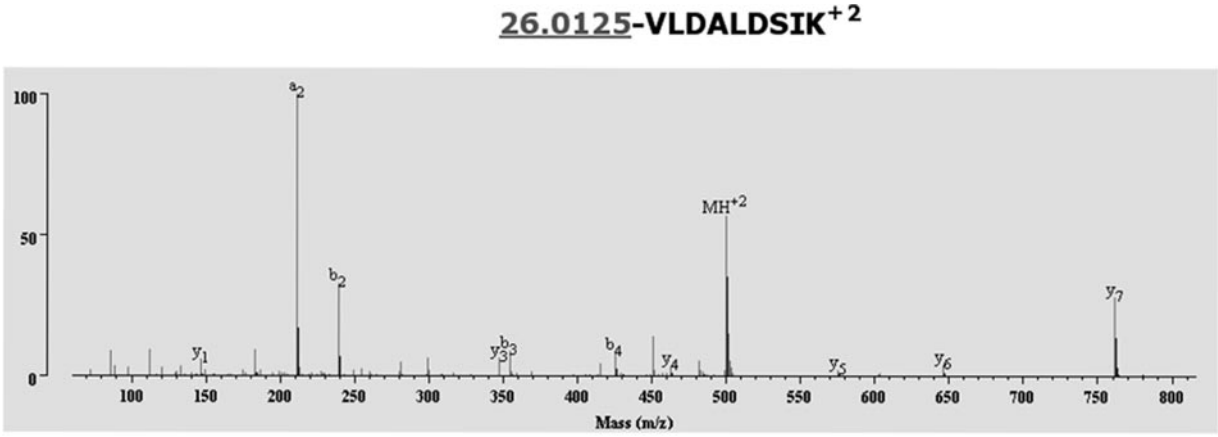


y2+	y		b	b2+
		V		
805.4139	1609.8205	P	197.1285	
756.8875	1512.7678	T	298.1761	
706.3637	1411.7201	P	395.2289	
657.8373	1314.6673	N	509.2718	
600.8158	1200.6244	V	608.3402	
551.2816	1101.5560	S	695.3723	
507.7656	1014.5239	V	794.4407	
458.2314	915.4555	V	893.5091	
408.6972	816.3871	D	1008.5360	
351.1837	701.3602	L	1121.6201	
294.6417	588.2761	T	1222.6678	
244.1179	487.2284	C(209.1003)	1534.7773	
88.0631	175.1190	R		

b



y2+	y		b	b2+
		V		
805.9904	1610.9735	P	197.1285	
757.4640	1513.9207	T	298.1761	
706.9401	1412.8730	P	395.2289	
658.4138	1315.8203	D	510.2558	
600.9003	1200.7933	V	609.3243	
551.3661	1101.7249	S	696.3563	
507.8501	1014.6929	V	795.4247	
458.3159	915.6245	V	894.4931	
408.7817	816.5560	D	1009.5201	
351.2682	701.5291	L	1122.6041	
294.7262	588.4450	T	1223.6518	
244.2023	487.3974	V	1322.7202	
194.6681	388.3289	K	1450.8152	725.9112
130.6206	260.2340	L	1563.8992	782.4533
74.0786	147.1499	A	1634.9363	817.9718
38.5600	76.1128	K(-71)		



y2+	y	26.0125-V	b
437.7477	874.4880	L	239.1723
381.2056	761.4040	D	354.1992
323.6921	646.3770	A	425.2363
288.1736	575.3399	L	538.3204
231.6316	462.2558	D	653.3473
174.1181	347.2289	S	740.3793
130.6021	260.1969	I	853.4634
74.0600	147.1128	K	

FIG. 7. **Identification of the peak list 2465 to the peptide VLDALDSIK with a mass modification of +26 Da on the peptide N terminus.** This modification corresponds to an acetaldehyde Schiff base modification to the N-terminal amine group.

modification to a peptide with a homologous sequence: VPTPDVSVVDLTVKLAK(−71). Fig. 6*b* displays this match and shows that the Inspect assignment does not match ions y1–y5 (the region where the two peptide assignments of Protein Prospector and Inspect differ most significantly). Also the mass accuracy of most of these peak matches is significantly outside the 0.1 Da mass accuracy expected. The 209 Da modification is formed by the alkylation of the Cys-SH with dithiothreitol (used to reduce disulfide bridges) followed by carbamidomethylation of the dithiothreitol on its other –SH group (caused by the addition of iodoacetamide to the mixture to alkylate free cysteine residues).

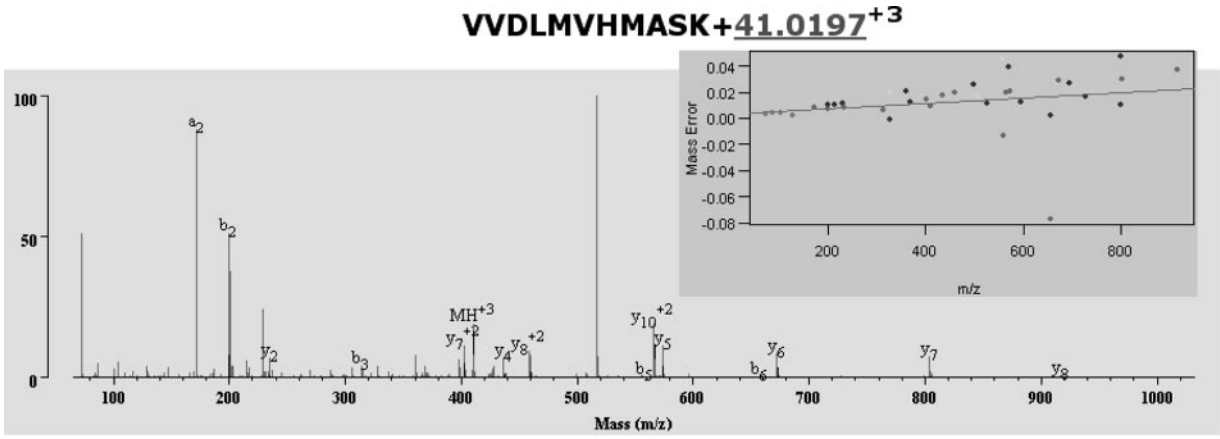
Another modification observed (13 times) is an addition of 26 Da. This has been reported previously as an acetaldehyde Schiff base modification to lysine residues (19), but we believe all the reported modifications in this analysis are to the peptide N terminus. An example spectrum of one of these modified peptides is shown in Fig. 7, which is peak list 2465. The modification can be restricted to one of the two most N-terminal residues by the b2 ion. In all the spectra detected with this modification, all b ions are always modified.

Protein Prospector also identified a few spectra that represented peptides with adducts. Spectrum 1480, shown in Fig. 8, is matched to an adduct of 41 Da. No modification of +41 Da is listed in Unimod, but a modification of this mass has been reported as an acetonitrile adduct in small molecule mass spectrometry studies (20). In the methods section of the publication accompanying this standard data set the authors state that the samples were stored in 1% acetonitrile (4). This modification appears to be eliminated as a neutral loss before the peptide backbone is fragmented, producing a fragmentation spectrum of an unmodified species. Protein Prospector reported four matches to adducts of 41 Da. Inspect did not detect this modification because it cannot look for modifications that cause neutral losses.

DISCUSSION

There are many mass spectrometry MSMS database search engines available for the research community. Hence it can be difficult to assess which search engine is the most appropriate choice for analyzing a given data set as comparisons of search engine performances are fraught with complications because of different searching parameters and thresholding for reporting the reliability of results. However, what is clear is that most search engines work better with certain types of data in preference to others. For example, it is widely recognized that Sequest, X!Tandem, and OMSSA are most effective with ion trap data. A goal of the Protein Pro-

Fig. 6. **Identification of the peak list 1975.** A precursor of *m/z* 854.931 was fragmented, and Protein Prospector reported a confident assignment to the peptide VPTPNVSVVDLTCR with a modification of +209 Da on the cysteine residue (*a*). Inspect reported a match to a homologous peptide, VPTPDVSVVDTVKLAK, with a modification of a loss of 71 Da from the C-terminal lysine (*b*). Plots in the *top right corners* of each panel show mass errors for fragment ion assignments. The intensity of the precursor ion in these figures has been artificially reduced to allow easier visualization of the fragment ions.



y2+	y		b	b2+
		V		
565.79	1130.57	V	199.14	
516.25	1031.50	D	314.17	
458.74	916.47	L	427.26	
402.20	803.39	M	558.30	
336.68	672.35	V	657.36	
287.14	573.28	H	794.42	397.72
218.61	436.22	M	925.46	463.24
153.09	305.18	A	996.50	498.75
117.58	234.14	S	1083.53	542.27
74.06	147.11	K		

FIG. 8. **Identification of peak list 1480 as VVDLMVHMASK with a modification of 41 Da somewhere on the peptide that is subsequently lost as a neutral group and therefore is not present on any of the fragment ions.** The plot in the *top right corner* shows the mass errors for fragment ion assignments.

spector software is to try to support data from all types of mass spectrometric platforms. The results presented here suggest that Protein Prospector has wider platform applicability than Sequest. The numbers of peptides identified in the non-tryptic searches allowing for extensive user-defined modifications show similar numbers of peptides identified in LTQ and 4800 data sets with the list of QSTAR-identified peptides being not much shorter. This contrasts with the Sequest results that suggested that the LTQ data set was significantly more information-rich than others. Protein Prospector uses slightly different scoring and looks for different ion types depending on the instrument geometry, and this extra flexibility makes it more generally applicable to a wider variety of instruments.

As this study is based on published standard data sets, hopefully other search engines will be evaluated using this data set in the future, so further reference points comparing search engines for different types of data can be obtained. It should also be pointed out that for purposes of direct comparison with the published Sequest results identical peak lists were used for searching. For the QSTAR data set charge states were not assigned during creation of the peak lists even though the data were good enough for charge state determination. Creating new peak lists with charge states assigned would halve the number of peak lists to be searched and would probably lead to halving the number of false positive identifications at a given acceptance threshold.

In the comparison of results from unexpected mass modification searching between Protein Prospector and Inspect, it appears that Protein Prospector performed better at reliably identifying more of the spectra. One factor contributing to this improvement was the ability of Protein Prospector to be able to consider a combination of user-defined modifications along with one unexpected mass modification. This contrasts with MS-Alignment of Inspect that can only consider unexpected mass modifications. The data could have been searched to allow for two mass modifications in Inspect, which would have potentially allowed Inspect to correctly match a few more spectra. However, in practice, any search allowing for two unspecified mass modifications produces results that are extremely unreliable. Thus, we would not recommend using this searching strategy.

The number of results reported with an expectation value less than 0.1 in the undefined mass modification search was similar to the reported number of matches in the search of the same data set with defined mass modifications. However, there are clearly many correct matches in the undefined mass modification search that cannot be in the other search results. This suggests there must be answers in the defined modification search that are no longer reported, or deemed confident, in the mass modification search. This is not very surprising given that the mass modification search is considering orders of magnitude more possibilities. This is exemplified in Table II, which lists the average number of precursors that were considered in the

TABLE II
Comparison of average number of precursors (search space) considered for the three types of searches of QSTAR data presented in this study

Search	Details	Average number of precursors considered
"Regular"	All Swiss-Prot, 4 considered modifications	14,859
"Extensive defined modifications"	46 proteins with 24 considered modifications	26,253
"Undefined mass modifications"	46 proteins with any mass modification between −100 and +300 Da	2,331,839

three types of search used on the QSTAR data set in this study. As can be seen, the initial search of the whole of Swiss-Prot database and the subsequent extensive variable modification search of the 46 proteins reported in the initial search are of somewhat comparable size and therefore speed with the modification search being slightly larger. However, the mass modification search is roughly 2 orders of magnitude larger, which, if everything else was equal, would mean expectation values for a given match in this search are going to be 2 orders of magnitude less confident than the same match in the variable modification or initial search. It also explains why the mass modification searching is more than 10 times slower. This highlights one of the issues of measuring the reliability of database searching in that Expectation values are a measure of the reliability assuming all possibilities are being considered and all are deemed equally possible, so if you consider many more precursor peptide options by allowing for more modifications, the reliability estimation becomes more conservative.

In both of these searches a large percentage of the spectra are being identified. There were a total of 1867 spectra acquired (which were duplicated as 2+ and 3+ precursors to produce 3734 peak lists). Hence well over half of the spectra are identified, and combining the two searching strategies the number is probably close to two-thirds of all spectra assigned. This is also despite the fact that precursors were only considered as doubly or triply charged, whereas in reality there were a few spectra of precursors at higher charge states. New peak lists were created in-house from these raw data that had correct precursor charge state assignments. The data were then searched allowing for variable modifications (not unknown mass modifications), and three spectra were confidently matched to 4+ precursors, and four spectra were confidently matched to 5+ precursors (data not shown). Despite this claim of generally better performance using Protein Prospector, we believe that searching using both softwares is a sensible option if one wants to fully characterize a sample. As can be seen in Fig. 2, a significant number of answers were confidently reported by Inspect that Protein Prospector agreed upon but did not deem significant matches. As the two softwares use significantly different approaches for identification, if both report the same result this does add weight to the assignment.

The emphasis of this presented work was to allow meaningful comparison of results between Protein Prospector and the published Sequest results rather than necessarily to analyze the published data set as comprehensively as possible. Hence the identical mzXML peak lists that were used for the Sequest analysis were also used in this study. In fact, these peak lists are not perfect representations of the raw data as charge states are not assigned to the precursor ions despite the data being unquestionably good enough to determine precursor ion charge states for some instruments. Also manually looking at some of the raw data, there are quite noticeable problems with labeling monoisotopic peak masses of multiply charged fragment peaks; the masses labeled quite often differ significantly from the observed peak, and information that would allow recognition of these peaks as multiply charged is often lost. Hence there is still clearly room for improvement in the analysis of these standard data sets.

There are several advantages of the Protein Prospector software over the alternatives for mass modification analysis. Its ability to perform searches allowing for a mixture of defined and unknown mass modifications gives greater flexibility. It can look for very large mass modifications, making it a powerful tool for identification of cross-linked peptides (12). It can identify modifications that are eliminated as neutral losses upon peptide fragmentation, such as sulfation and *O*-glycosylation, that alternative mass modification software cannot consider. It can access the raw data for many instrument formats, allowing identification of whether the monoisotopic peak was correctly labeled or if there was a co-eluting compound that may have been isolated and fragmented at the same time. It is part of the same search engine that does the conventional style database searching, so transfer of information from an initial search that produces the list of candidate proteins to this mass modification search is nothing more than the click of a button. This may seem a trivial advantage, but the effort required to take results from one program and submit them into another is generally sufficient to prevent people from trying this type of search. Finally as this software is part of a larger set of proteomics tools, there are direct links to other software that can be used to help verify results. This software is freely available on the web at http://prospector2.ucsf.edu/.

* This work was supported, in whole or in part, by National Institutes of Health Grant P41 RR001614. This work was also supported by the Vincent J. Coates Foundation. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

‡ To whom correspondence should be addressed: Dept. of Pharmaceutical Chemistry, University of California, 600 16th St., Genentech Hall, Rm. N474A, San Francisco, CA 94158-2517. E-mail: chalkley@cgl.ucsf.edu.

REFERENCES

1. Deutsch, E. W., Lam, H., and Aebersold, R. (2008) Data analysis and bioinformatics tools for tandem mass spectrometry in proteomics. *Physiol. Genomics* **33,** 18–25
2. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567
3. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20,** 1466–1467
4. Klimek, J., Eddes, J. S., Hohmann, L., Jackson, J., Peterson, A., Letarte, S., Gafken, P. R., Katz, J. E., Mallick, P., Lee, H., Schmidt, A., Ossola, R., Eng, J. K., Aebersold, R., and Martin, D. B. (2008) The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. *J. Proteome Res.* **7,** 96–103
5. Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4,** 207–214
6. Nielsen, M. L., Savitski, M. M., and Zubarev, R. A. (2006) Extent of modifications in human proteome samples and their effect on dynamic range of analysis in shotgun proteomics. *Mol. Cell. Proteomics* **5,** 2384–2391
7. Tsur, D., Tanner, S., Zandi, E., Bafna, V., and Pevzner, P. A. (2005) Identification of post-translational modifications by blind search of mass spectra. *Nat. Biotechnol.* **23,** 1562–1567
8. Savitski, M. M., Nielsen, M. L., and Zubarev, R. A. (2006) ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures. *Mol. Cell. Proteomics* **5,** 935–948
9. Han, Y., Ma, B., and Zhang, K. (2005) SPIDER: software for protein identification from sequence tags with de novo sequencing error. *J. Bioinform. Comput. Biol.* **3,** 697–716
10. Searle, B. C., Dasari, S., Turner, M., Reddy, A. P., Choi, D., Wilmarth, P. A., McCormack, A. L., David, L. L., and Nagalla, S. R. (2004) High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal. Chem.* **76,** 2220–2230
11. Baker, P. R., Chalkley, R. J., Medzhiradszky, K. F., Snedecor, J. O., and Burlingame, A. L. (2006) Improved methods for comprehensive sample analysis using Protein Prospector, in *Proceedings of the 54th ASMS Conference on Mass Spectrometry and Allied Topics, Seattle, May 28–June 1, 2006*, TP25 44, American Society for Mass Spectrometry, Santa Fe, NM
12. Chalkley, R. J., Baker, P. R., Medzhiradszky, K. F., and Burlingame, A. L. (2007) Discovery of unanticipated modifications using Protein Prospector, in *55th ASMS Conference on Mass Spectrometry, Indianapolis, June 3–7, 2007*, MPI 189, American Society for Mass Spectrometry, Santa Fe, NM
13. Chalkley, R. J., Baker, P. R., Huang, L., Hansen, K. C., Allen, N. P., Rexach, M., and Burlingame, A. L. (2005) Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets. *Mol. Cell. Proteomics* **4,** 1194–1204
14. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989
15. Fenyo, D., and Beavis, R. C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75,** 768–774
16. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumby, M., Pevzner, P. A., and Bafna, V. (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **77,** 4626–4639
17. Beausoleil, S. A., Villen, J., Gerber, S. A., Rush, J., and Gygi, S. P. (2006) A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24,** 1285–1292
18. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., and Mann, M. (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–648
19. Braun, K. P., Cody, R. B., Jr., Jones, D. R., and Peterson, C. M. (1995) A structural assignment for a stable acetaldehyde-lysine adduct. *J. Biol. Chem.* **270,** 11263–11266
20. D'Agostino, P. A., Hancock, J. R., and Provost, L. R. (1999) Packed capillary liquid chromatography-electrospray mass spectrometry analysis of organophosphorus chemical warfare agents. *J. Chromatogr. A* **840,** 289–294
21. Creasy, D. M., and Cottrell, J. S. (2004) Unimod: protein modifications for mass spectrometry. *Proteomics* **4,** 1534–1536